TOKENIZER FOR A NATURAL LANGUAGE PROCESSING SYSTEM

ABSTRACT OF THE DISCLOSURE

The present invention is a segmenter used in a natural language processing system. The segmenter segments a textual input string into tokens for further natural language processing. In accordance with one feature of the invention, the segmenter includes a tokenizer engine that proposes segmentations and submits them to a linguistic knowledge component for validation. In accordance with another feature of the invention, the segmentation system includes language-specific data that contains a precedence hierarchy for punctuation. If proposed tokens in the input string contain punctuation, they can illustratively be broken into subtokens based on the precedence hierarchy.